Assisted Curation

Alex, B., Grover, C., Haddow, B., Kabadjor, M., Klein, E., Matthews, M., Roebuck, S., Tobin, R., & Wang, X. (2008). Assisted Curation: Does Text Mining Really Help?. In R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, & T. E. Klein (Eds.), Proceedings of the Pacific Symposium on Biocomputing (Biocomputing 2008). (pp. 556-567). Singapore: World Scientific Press.

ASSISTED CURATION: DOES TEXT MINING REALLY HELP?

BEATRICE ALEX, CLAIRE GROVER, BARRY HADDOW, MIJAIL KABADJOV, EWAN KLEIN, MICHAEL MATTHEWS, STUART ROEBUCK, RICHARD TOBIN, AND XINGLONG WANG

School of Informatics
University of Edinburgh
EH8 9LW, UK
E-mail for correspondence: balex@inf.ed.ac.uk

Although text mining shows considerable promise as a tool for supporting the curation of biomedical text, there is little concrete evidence as to its effectiveness. We report on three experiments measuring the extent to which curation can be speeded up with assistance from Natural Language Processing (NLP), together with subjective feedback from curators on the usability of a curation tool that integrates NLP hypotheses for protein-protein interactions (PPIs). In our curation scenario, we found that a maximum speed-up of 1/3 in curation time can be expected if NLP output is perfectly accurate. The preference of one curator for consistent NLP output and output with high recall needs to be confirmed in a larger study with several curators.

1. Introduction

Curating biomedical literature into relational databases is a laborious task requiring considerable expertise, and it is proposed that text mining should make the task easier and less time-consuming [1, 2, 3]. However, to date, most research in this area has focused on developing objective performance metrics for comparing different text mining systems (see [4] for a recent example). In this paper, we describe initial feedback from the use of text mining within a commercial curation effort, and report on experiments to evaluate how well our NLP system helps curators in their task.

This paper is organised as follows. We review related work in Section 2. In Section 3, we introduce the concept of assisted curation and describe the different aspects involved in this process. Section 4 provides an overview of the components of our text mining system, the TXM (text mining) NLP pipeline, and describes the annotated corpus used to train and evaluate this system. In Section 5, we describe and discuss the results of three different curation experiments which attempt to test the effectiveness of various versions of the NLP pipeline in assisting curation. Discussion and conclusions follow in Section 6.

2. Related Work

Despite the recent surge in the development of information extraction (IE) systems for automatic curation of biomedical data spurred on by the BioCreAtIvE II competition [5], there is a lack of user studies that extrinsically evaluate the usefulness of IE as a way to assist curation. Donaldson et al. [6] reported an estimated 70% reduction in curation time of yeast-protein interactions when using the PreBIND/Textomy IE system, designed to recognise abstracts containing protein interactions. This estimate is limited to the document selection component of PreBind and does not include time savings due to automatic extraction and normalization of named entities (NEs) and relations. Karamanis et al. [7] studied the functionality and usefulness of their curation tool, ensuring that integrating NLP output does not impede curators in their work. In three curation experiments with one curator, they found evidence that improving their curation tool and integrating NLP speeds up curation compared to using a tool prototype with which the curator was not experienced at the start of the experiment. Karamanis et al. [7] mainly focus on tool functionality and presentational issues. They did not analyse the aspects of the NLP output that were useful to curators, how it affected their work, or how the NLP pipeline can be tuned to simplify the curator's job. Recently, Hearst et al. [8] reported on a pilot usability study showing positive reactions to figure display and caption search for bioscience journal search interfaces.

Regarding non-biomedical-related applications, Kristjansson et al. [9] describe an interactive IE tool with constraint propagation to reduce human effort in address form filling. They show that highlighting contact details in unstructured text, pre-populating form fields, and interactive error correction by the user reduces the cognitive load on users when entering address details into a database. This reduction is reflected in the expected number of user actions, which is determined based on the number of clicks to enter all fields. They also integrated confidence values to inform the user about the reliability of extracted information.

3. Assisted Curation

The curation task that we will discuss in this paper requires curators to identify examples of protein-protein interactions (PPIs) in biomedical literature. The initial step involves retrieving a set of papers that match criteria for the curation domain. After an initial step of further filtering the papers into promising candidates for curation, curators proceed on a paper-by-paper basis. Using an inhouse editing and verification tool (henceforth referred to as the 'Editor'), the curators are able to read through an electronic version of the paper and enter retrieved information into a template which will then be used to add a record to a relational database.

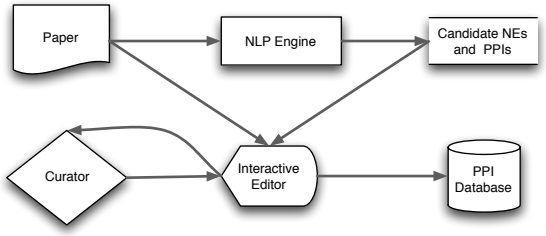

Figure 1. Information Flow in the Curation Process

Curation is a laborious task which requires considerable expertise. The curator spends a significant amount of time on reading through a paper and trying to locate material that might contain curatable facts. Can NLP help the curator work more efficiently? Our basic assumption, which is commonly held [1], is that IE techniques are likely to be effective in identifying relevant entities and relations. More specifically, we assume that NLP can propose *candidate* PPIs; if the curators restrict their attention to these candidates, then the time required to explore the paper can be reduced. Notice that we are not proposing that NLP should *replace* human curators—given the current state of the art, only expert humans can assure that the captured data is of sufficiently high quality to be entered into databases.

Our curation scenario is illustrated in Figure 1. The source paper undergoes processing by the NLP engine. The result is a set of normalised NEs and candidate PPIs. The original paper and the NLP output are fed into the interactive Editor, which then displays a view to the curator. The curator makes a decision about which information to enter into the Editor, which is then communicated to a backend database.

In one sense, we can see this scenario as one in which the software provides *decision support* to the human. Although in broad terms the decision is about what facts, if any, to curate, this can be broken down into smaller subtasks. Given a sentence S, (i) do the terms in S name proteins? If so, (ii) which proteins do they name? And (iii), given two protein mentions, do the proteins stand in an interaction relation? These decision subtasks correspond to three components of the NLP engine: (i) Named Entity Recognition, (ii) Term Identification, and (iii) Relation Extraction. We will examine each of these in turn shortly, but first, we want to consider further the kind of choices that need to be made in examining the usability of NLP for curation. A crucial observation is that the NLP output is bound to be imperfect. How can the curator make use of an *unreliable* assistant?

First, there are **interface design** issues—what information is displayed to the curator, in what form, and what kind of manipulations can the curator carry out?

Second, what is the **division of labour** between the human and the software? For example, there might be some decisions which are relatively cheap for the curator to make, such as deciding what species is associated with a protein mention, and which can then help the software in providing a more focused set of candidates for term identification.

Third, what are the optimal **functional characteristics** of the NLP engine, given that complete reliability is not currently attainable? For example, should the NLP try to improve recall over precision, or *vice versa*?

Although the first and second dimensions are clearly important, in this paper we will focus on the third, namely the functional characteristics of our system.

4. TXM Pipeline

The NLP output displayed in the interactive curation Editor is produced by the TXM pipeline, an IE pipeline that is being developed for use in biomedical IE tasks. The particular version of the pipeline used in the experiments described here focuses on extracting proteins, their interactions, and other entities which are used to enrich the interactions with extra information of biomedical interest. Proteins are also normalised (i.e., mapped to identifiers in an appropriate database) using the term identification (TI) component of the pipeline. In this section a brief description of the pipeline, and the corpus used to develop and test it, will be given, with more implementation details provided by appropriate references.

Corpus In order to use machine learning approaches for named entity recognition (NER) and relation extraction (RE), and for evaluating the pipeline components, an annotated corpus was produced using a team of domain experts. Since the annotations contain information about proteins and their interactions, it is referred to as the enriched protein-protein interaction (EPPI) corpus. The corpus consists of 217 full-text papers selected from PubMed and PubMedCentral as containing experimentally proven PPIs. The papers, retrieved in XML or HTML, were converted to an internal XML format. Nine types of entities (Complex, CellLine, DrugCompound, ExperimentalMethod, Fusion, Fragment, Modification, Mutant, and Protein) were annotated, as well as PPI relations and FRAG relations (which link Fragments or Mutants to their parent proteins). Furthermore, proteins were normalised to their RefSeq[a] identifier and PPIs were enriched with properties and attributes. The properties added to the PPIs are IsProven, IsDirect and IsPositive and the possible attributes are CellLine, DrugTreatment, ExperimentalMethod or ModificationType. More details on properties and attributes can be found in Haddow

[a] http://www.ncbi.nlm.nih.gov/RefSeq/index.html

and Matthews [10]. The inter-annotator agreement (IAA), measured on a sample of doubly and triply annotated papers, amounts to an overall micro-averaged F1-score[b] of 84.9 for NEs, 88.4 for normalisations, 64.8 for PPI relations, 87.1 for properties and 59.6 for attributes. The EPPI corpus ($\backsimeq$2m tokens) is divided into three sections, TRAIN (66%), DEVTEST (17%), and TEST (17%).

Pre-processing A set of pre-processing steps in the pipeline was implemented using the LT-XML2 tools [11]. The pre-processing performs sentence boundary detection and tokenization, adds useful linguistic markup such as chunks, part-of-speech tags, lemmas, verb stems, and abbreviation information, and also attaches NCBI taxonomy identifiers to any species-related terms.

Named Entity Recognition The NER component is based on the C&C tagger, a Maximum Entropy Markov Model (MEMM) tagger developed by Curran and Clark [12], and augmented with extra features and gazetteers tailored to the domain and described fully in Alex et al. [13]. The C&C tagger allows for the adjustment of the entity decision threshold through the `prior` file, which has the effect of varying the precision-recall balance in the output of the component. This `prior` file was modified to produce the high precision and high recall models used in the assisted curation experiment described in Section 5.3.

Term Identification The TI component uses a rule-based fuzzy matcher to produce a set of candidate identifiers for each recognized protein. Species are assigned to proteins using a machine learning based tagger trained on contextual and species word features [14]. The species information and a set of heuristics are used to choose the most probable identifiers from the set of candidates proposed by the matcher. The evaluation metric for the TI system is *bag accuracy*. This means that if the system produces multiple identifiers for an entity mention, it is counted as a hit as long as one of the identifiers is correct. The rationale is that since a TI system that outputs one identifier is not accurate enough, generating a bag of choices increases chances of finding the correct one. This can assist curators as the right identifier can be chosen from a bag (see [15] for more details).

Relation Extraction Intra-sentential PPI and FRAG relations are both extracted using the system described in Nielsen [16], with inter-sentential FRAG relations addressed using a maximum entropy model trained on features derived from the entities, their context, and other entities in the vicinity. Enriching the relations with properties and attributes is implemented using a mixture of machine learning and rule-based methods described in Haddow and Matthews [10].

[b] Micro-averaged F1-score means that each example is given equal weight in the evaluation.

Component Performance The performance of the IE components of the pipeline (NER, TI, and RE) is measured using precision, recall, and F1-score (except TI– see above), by testing each component in isolation and comparing its output to the annotated data. For example, RE is tested using the annotated (gold) entities as its input, rather than the output of NER, in order that NER errors not affect the score for RE. Table 1 shows the performance of each component when tested on DEVTEST , where the machine learning components are trained on TRAIN .

Table 1. Performance of pipeline components, tested in isolation on DEVTEST and trained on TRAIN.

Component	TP	FP	FN	Precision	Recall	F1
NER (micro-average)	19,925	5,964	7,755	76.96	71.98	74.39
RE (PPI)	1,208	1,173	1,080	50.73	52.80	51.75
RE (FRAG)	1,699	963	1,466	63.82	53.68	58.31
RE (properties micro-average)	3,041	567	579	84.28	84.01	84.14
RE (attributes micro-average)	483	822	327	37.01	59.63	45.67
Component	TP	FP	FN	Precision	Recall	Bag Acc.
TI (micro-average)	9,078	91,396	2,843	9.04	76.15	76.15

5. Curation Experiments

We conducted three curation experiments with and without assistance from the output of the NLP pipeline or gold standard annotations (GSA). In all of the experiments, curators were asked to curate several documents according to internal guidelines. Each paper is assigned a curation ID for which curators create several records corresponding to the curatable information in the document. Curators always use an interactive Editor which allows them to see the document on screen and enter the curatable information into record forms. All curators are experienced in using the interactive curation Editor, but not necessarily familiar with assisted curation. After completing the curation for each paper, they were asked to fill in a questionnaire.

5.1. *Manual versus Assisted Curation*

In the first experiment, 4 curators curated 4 papers in 3 different conditions:

- **MANUAL:** without assistance
- **GSA-assisted:** with integrated gold standard annotations
- **NLP-assisted:** with integrated NLP pipeline output

Each curator processed a paper only once, in one specific condition, without being informed about the type of assistance (GSA or NLP), if any. This experiment

Table 2. Total number of records curated in each condition and average curation speed per record.

		Time per record	
Condition	Records	Average	StDev
MANUAL	121	312s	327s
GSA	170	205s	52s
NLP	141	243s	36s

Table 3. Average questionnaire scores. Scores ranged from (1) for strongly agree to (5) for strongly disagree.

Statement	GSA	NLP
NLP was helpful in curating this documents	2.75	3.25
NLP speeded up the curation of this paper	3.75	3.75
NE annotations were useful for curation	2.50	3.00
Normalizations of NEs were useful for curation	2.75	2.75
PPIs were useful for curation	3.50	3.25

aims to answer the following questions: Does the NLP output which is currently integrated in the interactive Editor accelerate curation? Secondly, do human gold standard annotations assist curators in their work—i.e., how helpful would NLP be to a curator if it performed as well as a human annotator?

Table 2 shows that for all four papers, the fewest records (121) were curated during manual curation, 20 more records (+16.5%) were curated given NLP assistance, and 49 more records (+40.5%) with GSA assistance. This indicates that providing NLP output helps curators to spot more information. Ongoing work involves a senior curator assessing each curated record in terms of quality and coverage. This will provide evidence for whether this additional information is also curatable, i.e. how the NLP output affects curation accuracy, and also give an idea of inter-curator agreement for different conditions. As each curator curated in all three conditions but never curated the same paper twice, inter-document and inter-curator variability must be considered. Therefore, we present curation speed per condition as the average speed of curating a record. Manual curation is most time-consuming, followed by NLP-assisted curation (22% faster), followed by GSA-assisted curation (34% faster). Assisted curation clearly speeds up the work of a curator, and a maximum reduction of $1/3$ in manual curation time can be expected if the NLP pipeline performed with perfect accuracy.

In the questionnaire, curators rated GSA assistance slightly more positively than NLP assistance (see Table 3). However, they were not convinced of either condition speeding up their work, even though the time measurements show otherwise. Considering that they were not familiar with assisted curation prior to the experiment, a certain effect of learning should be allowed for. Moreover, they

Table 4. Total number of records curated in each consistency condition and average curation speed per record.

	Time per record	
Condition	Average	StDev
CONSISTENCY1	128s	43s
CONSISTENCY2	92s	22s

may have had relatively high expectations of the NLP output. In fact, individual feedback in the questionnaire shows that NLP assistance was useful for some papers and some curators, but not others. Further feedback in the questionnaire includes aspects of visualization (e.g. PDF conversion errors) and interface design (e.g. inadequate display of information linked to NE normalizations) in the interactive Editor. Regarding the NLP output, curators also requested more accurate identification of PPI candidates, e.g. in coordinations like "A and B interact with C and D", and more consistency in the NLP output.

5.2. *NLP Consistency*

The NLP pipeline extracts information based on context features and may, for example, recognize a string as a protein in one part of the document but as a drug/compound in another, or assign different species to the same protein mentioned multiple times in the document. While this inconsistency may not be erroneous, the curators' feedback is that consistency would be preferred. To test this hypothesis, and to determine whether consistent NLP output helps to speed up curation, we conducted a second experiment. One curator was asked to curate 10 papers containing NLP output made consistent in two ways. In 5 papers, all NEs recognized by the pipeline were propagated throughout the document (CONSISTENCY1). In the other 5 papers, only the most frequent NE recognized for a particular surface form is propagated, while less frequent ones are removed (CONSISTENCY2). In both conditions, the most frequent protein identifier bag determined by the TI component is propagated for each surface form, and ePPIs are extracted as usual. Subsequent to completing the questionnaire, the curator viewed a second version of the paper in which consistency in the NLP output was not forced, and filled in a second questionnaire regarding the comparison of both versions.

Table 4 shows that the curator managed to curate 28% faster given the second type of consistency. However, examining the answers to the questionnaire listed in Table 5, it appears that the curator actually considerably preferred the first type of consistency, where all NEs that were recognized by the NER component are propagated throughout the paper. While this speed-up in curation may be attrac-

Table 5. Average questionnaire scores. Scores ranged from (1) for strongly agree to (5) for strongly disagree. In questionnaire 2, consistent (CONSISTENCY1/2) NLP output (A) is compared to baseline NLP (B).

Statement	CONSISTENCY1	CONSISTENCY2
Questionnaire 1		
NLP output was helpful for curation	1.6	2.6
NLP output speeded up curation	1.8	3.2
NEs were useful for curation	1.4	4.0
Normalizations of NEs were useful for curation	3.2	4.0
PPIs were useful for curation	3.6	4.2
Questionnaire 2		
A was more useful for curation than B would have been	2.6	4.0
A speeded up the curation process more than B would have	3.0	4.0
A appeared more accurate than B	2.6	4.2
A missed important information compared to B	4.4	1.8
A contained too much information compared to B	3.6	4.6

tive from a commercial perspective, this experiment illustrates how important it is to get feedback from users who may well reject a technology altogether if they are not happy working with it.

5.3. *Optimizing for Precision or Recall*

Currently, all pipeline components are optimized for F1-score, resulting in a relative balance between the correctness and coverage of extracted information, i.e. precision and recall. In previous curation rounds, curators felt they could not completely trust the NLP output, as some of the information displayed was incorrect. The final curation experiment tests whether optimizing the NLP pipeline for F1 is ideal in assisted curation, or whether a system that is more correct but misses some curatable information (high precision) or one that extracts most of the curatable information along with many non-curatable or incorrect facts (high recall) would be preferred. In this experiment, only the NE component was adapted to increase its precision or recall. This is done by changing the threshold in the C&C `prior` file to modify tag probabilities assigned by the C&C tagger.[c] The intrinsic evaluation scores of the NER component optimized either for F1, precision, or recall are listed in Table 6.

In the experiment, one curator processed 10 papers in random order containing NLP output, 5 with high recall NER and 5 with high precision. Note that to simplify

[c]Internal and external features were not optimized for precision or recall. This could be done to increase effects even more. The TI and RE components were also not modified for this experiment.

Table 6. Optimized F1-score versus high precision (P) and high recall (R) NER, along with corresponding counts of true positives (TP), false positives (FP), and false negatives (FN).

Setting	TP	FP	FN	P	R	F1
High F1	20,091	6,085	7,589	76.75	72.58	**74.61**
High P	11,836	1,511	15,844	**88.68**	42.76	57.70
High R	21,880	20,653	5,800	51.44	**79.05**	62.32

the experiment the curator did not normalise entities in this curation round. Subsequent to completing the questionnaire, the curator viewed a second version of the paper with NLP output based on optimized F1-score NER and filled in a second questionnaire regarding the comparison of both versions. The results in Table 7 show that the curator rated all aspects of the high recall NER condition more positively than of the high precision NER condition. Moreover, the curator tended to prefer NLP output with optimised F1 NER over that containing high precision NER, and NLP output containing high recall NER over that with high F1 NER. Although the number of curated papers is small, this curator seems to prefer NLP output that captures more curatable information but is overall less accurate. The curator noted that since her curation style involves skim-reading, the NLP output helped her to spot information that she otherwise would have missed. The results of this experiment could therefore be explained simply by curation style. Another curator with a more meticulous reading style may actually prefer more precise and trustworthy information extracted by the NLP pipeline. Clearly, the last curation experiment needs to be repeated using several curators, curating a larger set of papers, and providing additional timing information per curated record. In general, it would be useful to develop a system that will allow curators to filter information presented onscreen dynamically, possibly based on confidence values, as integrated in the tool described by Kristjansson et al. [9].

6. Discussion and Conclusions

This paper has focused on optimizing functional characteristics of an NLP pipeline for assisted curation, given that current text mining techniques for biomedical IE are not completely reliable. Starting with the hypothesis that assisted curation can support the task of a curator, we found that a maximum reduction of $1/3$ in curation time can be expected if NLP output is perfectly accurate. This shows that biomedical text mining can assist in curation. Moreover, NLP assistance led to the curation of more records, although the validity of this additional information still needs to be confirmed by a senior curator.

In extrinsic evaluation of the NLP pipeline in curation, we have tested several optimizations of the output in order to determine the type of assistance that is

Table 7. Average questionnaire scores. Scores ranged from (1) for strongly agree to (5) for strongly disagree. In questionnaire 2, optimized precision/recall (HighP/HighR) NER output (A) is compared to optimized F1 NER output (B).

Statement	HighP NER	HighR NER
Questionnaire 1		
NLP output was helpful for curation	3.0	2.2
NLP output speeded up curation	3.4	2.4
NEs were useful for curation	3.0	2.0
PPIs were useful for curation	3.2	2.5
Questionnaire 2		
A was more useful for curation than B would have been	4.2	2.6
A speeded up the curation process more than B would have	4.2	3.0
A appeared more accurate than B	4.4	2.8
A missed important information compared to B	1.4	3.2
A contained too much information compared to B	4.8	3.8

preferred by curators. We found that the curator prefers consistency, with all NEs propagated throughout the document, even though this preference is not reflected in the average time measurements for curating a record. When comparing curation with NLP output containing high recall or high precision NE predictions, the curator clearly preferred the former. While this result illustrates that optimizing an IE system for F1-score does not necessarily result in optimal performance in assisted curation, this experiment must be repeated with several curators in view of different curation styles.

Overall, we learnt that measuring curation in terms of curation time is not sufficient to capture the usefulness of NLP output for assisted curation. As recognized by Karamanis et al. [7], it is difficult to measure a curator's performance as one quantitative metric. The average time to curate a record, alone, is clearly not sufficient for capturing all factors involved the curation process. It is important to work closely with the user of a curation system in order to identify helpful and hindering aspects of such technology. In future work, we will conduct further curation experiments to determine the merit of high recall and high precision NLP output for the curation task. We will also invest some time in implementing confidence values of extracted information into the interactive Editor.

Acknowledgements

This work was carried out as part of an ITI Life Sciences Scotland (`http://www.itilifesciences.com`) research programme with Cognia EU (`http://www.cognia.com`) and the University of Edinburgh. The authors are very grateful to the curators at Cognia EU who participated in the experiments. The inhouse curation tool used for this work is the subject of International Patent Application No. PCT/GB2007/001170.

References

1. A. S. Yeh, L. Hirschman, and A. Morgan. Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. *Bioinformatics*, 19(Suppl 1):i331–339, 2003.
2. D. Rebholz-Schuhmann, H. Kirsch, and F. Couto. Facts from text - is text mining ready to deliver? *PLoS Biology*, 3(2), 2005.
3. H. Xu, D. Krupke, J. Blake, and C. Friedman. A natural language processing (NLP) tool to assist in the curation of the laboratory mouse tumor biology database. *Proceedings of the AMIA 2006 Annual Symposium*, page 1150, 2006.
4. L. Hirschman, M. Krallinger, and A. Valencia, editors. *Second BioCreative Challenge Evaluation Workshop*. Fundación CNIO Carlos III, Madrid, Spain, 2007.
5. M. Krallinger, F. Leitner, and A. Valencia. Assessment of the second BioCreative PPI task: Automatic extraction of protein-protein interactions. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 41–54, Madrid, Spain, 2007.
6. I. Donaldson, J. Martin, B. de Bruijn, C. Wolting, V. Lay, B. Tuekam, S. Zhang, B. Baskin, G.D. Bader, K. Michalickova, T. Pawson, and C.W.V. Hogue. PreBIND and Textomy - mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, 4 (11), 2003.
7. N. Karamanis, I. Lewin, R. Seal, R. Drysdale, and E. Briscoe. Integrating natural language processing with FlyBase curation. In *Proceedings of PSB 2007*, pages 245–256, Maui, Hawaii, 2007.
8. M. A. Hearst, A. Divoli, J. Ye, and M. A. Wooldridge. Exploring the efficacy of caption search for bioscience journal search interfaces. In *Proceedings of BioNLP 2007*, pages 73–80, Prague, Czech Republic, 2007.
9. T. T. Kristjansson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In Deborah L. McGuinness and George Ferguson, editors, *Proceedings of AAAI 2004*, pages 412–418, San Jose, US, 2004.
10. Barry Haddow and Michael Matthews. The extraction of enriched protein-protein interactions from biomedical text. In *Proceedings of BioNLP*, pages 145–152, Prague, Czech Republic, 2007.
11. C. Grover and R. Tobin. Rule-based chunking and reusability. In *Proceedings of LREC 2006*, pages 873–878, Genoa, Italy, 2006.
12. J. Curran and S. Clark. Language independent NER using a maximum en-

tropy tagger. In *Proceedings of CoNLL-2003*, pages 164–167, Edmonton, Canada, 2003.
13. B. Alex, B. Haddow, and C. Grover. Recognising nested named entities in biomedical text. In *Proceedings of BioNLP 2007*, pages 65–72, Prague, Czech Republic, 2007.
14. X. Wang. Rule-based protein term identification with help from automatic species tagging. In *Proceedings of CICLING 2007*, pages 288–298, Mexico City, Mexico, 2007.
15. X. Wang and M. Matthews. Comparing usability of matching techniques for normalising biomedical named entities. In *Proceedings of PSB 2008*, 2008.
16. L. A. Nielsen. Extracting protein-protein interactions using simple contextual features. In *Proceedings of BioNLP 2006*, pages 120–121, New York, US, 2006.